SMP, NUMA and Path to HPC

Name: Leonard Tsai

Title: Chief Technologist

Company: NEC

San Jose January 23-24, 2001 • Taipei February 14-15, 2001

Outline

- Path to HPC
- NEC Azusa overview
- Challenges
- Future
- Q & A

Outline

- **Path to HPC**
- NEC Azusa overview
- Challenges
- Future
- Q & A

Quest for HPC

- Major approaches:
  - Improvement on Micro architecture
  - Soc for multi process unit
  - SMP
  - MPP
  - Clustering
  - Peer to Peer

Micro Architecture

- Design Complexity and compromise
  - Multiple issue, out of order execution, super pipeline, super scalar, SIMD unit, …
  - Pentium 4 micro burst engine and hyper pipeline
  - Takes 5 to 7 years to architect next version
- Legacy Software pressure
  - Cost, time and resource for new software
  - Programmer generation and paradigm shift
- IA 64 is a major branch point with *EPIC*
  - It may take 5 to 7 years before catching up

Soc with Multi PU

- Extending existing Architecture
  - Either symmetric PU or special function PU
- Working silicon in market today
  - E.g. TI C64x, Sun Magc 5200 (announce in 2 weeks), Sony PS/2, …
- Coherent at Cache level
  - Most share common L2
- Specialize in MISD application
  - Network processor, Telcomm, packet processor, etc.

SMP

- Multiple CPUs in one system
  - Looks like single CPU, memory and I/O
  - Every CPU looks the same (symmetric)
- System Design Complexity
  - Bus vs. point to point
  - 80-20 rule
  - API and OS concern
  - Major system level design trend in last 20 years

MPP

- Extension of SMP
  - Only used in specialized market
  - Taking over vector Super computer market – see www.top500.org 90% are scalar, 10% are vector
- Niche market
  - IBM dominate the market with Sun, SGI, Cray, NEC, …
  - Specialized everything, including service engineers
  - Solve largely numerical intensive, time sensitive problem

Clustering

- System level MP approach
  - Using interlink to perform task
  - Exposed vs Enclosed cluster
  - Message based vs Share based
  - Multi-level clustering
- Commercial solution available and accepted
  - HA (High Availability)
  - FT (Fault Tolerant)
  - FR (Fault Resilient)
  - Load sharing and balance

Peer to Peer

- Distributed Computing
  - Loosely coupled, equal nodes processing same problem
  - Large, unstructured network problem
  - Centralize depository vs progressive mesh
  - Experiments & development sponsored by government – GRID, Condor, …
- Challenge
  - OS, client, and programming model
  - Overhead, cost, time, and other factors
  - Commercialization

Outline

- Path to HPC
- **NEC Azusa overview**
- Challenges
- Future
- Q & A

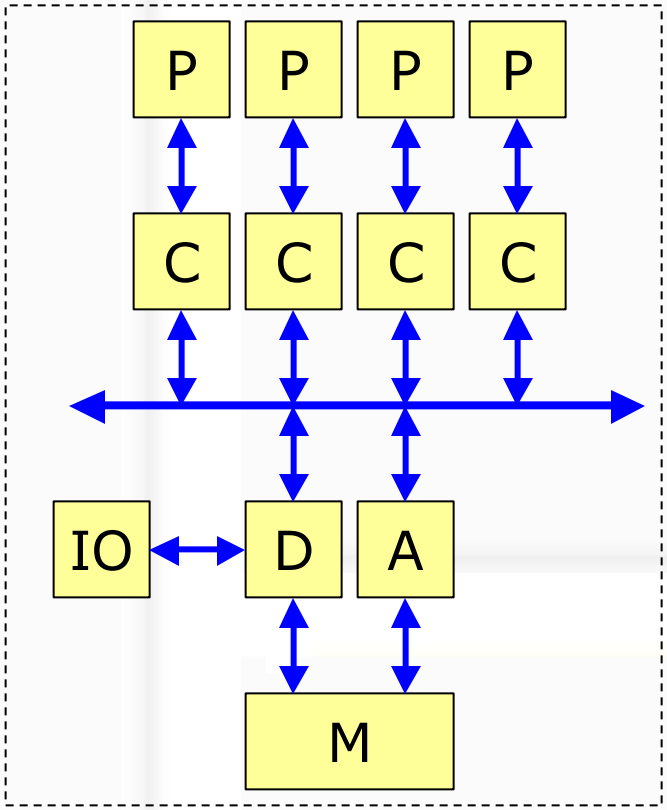
NEC Azusa

- 16 way Itanium Server
  - 4 years in development with 1000+ engineers
  - Functioning unit demo many times and for development
  - ccNUMA based SMP with 4 Quad cell; each 4 Itanium, 32 GB RAM for total of 128 GB
  - Enterprise level RSA features for mission critical application
  - Available when Intel ship Itanium CPU

Itanium QUAD Cell

- Four (4) processor units (with power pods) connected on a common bus to One (1) System Address Controller (A) and One (1) System Data Controller (D)
- One (1) common memory
- One (1) set of IO Controller (IOC) connect to the SAC and SDC

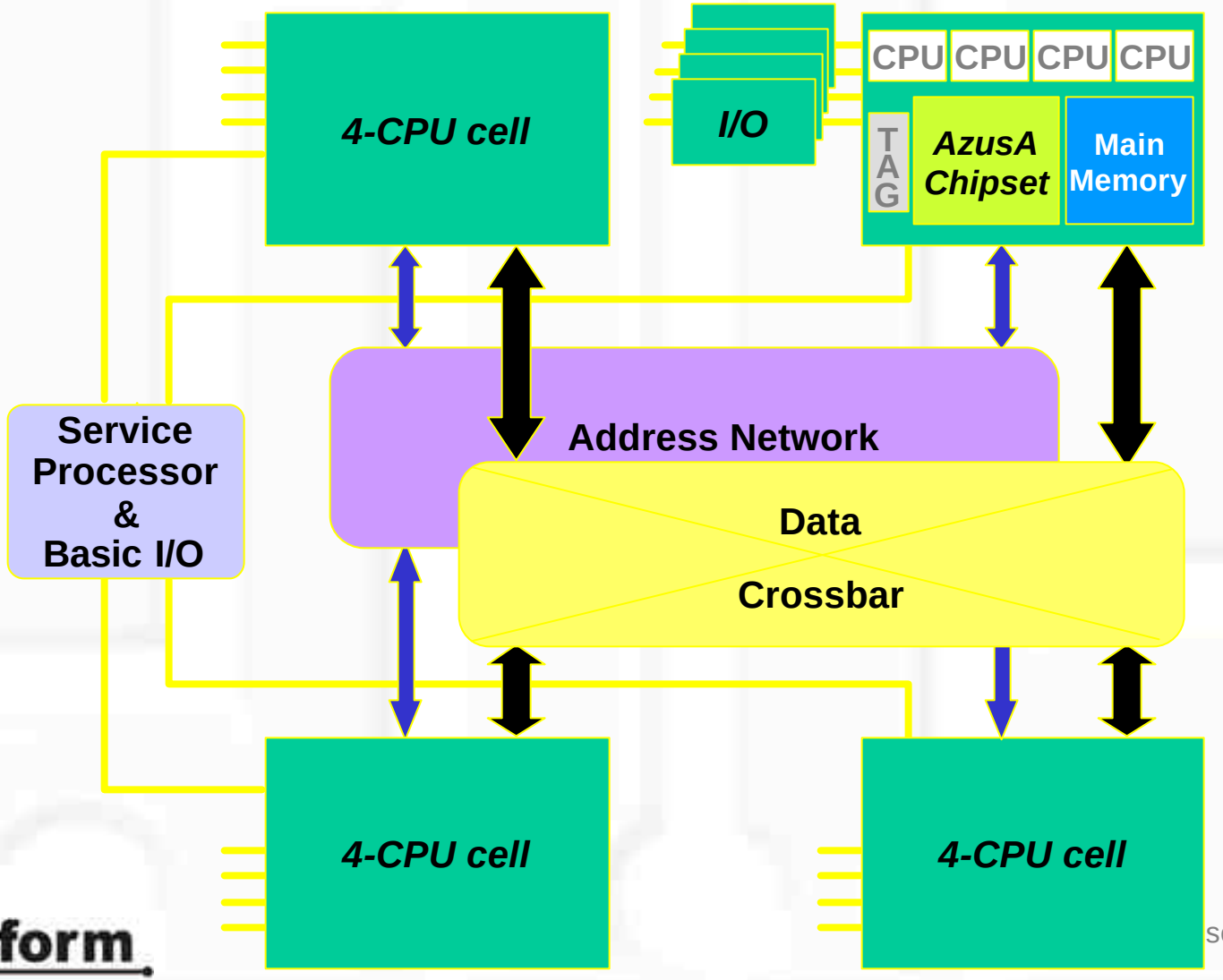
* Similar to Intel 860 approach



NEC Azusa Features

- Azusa uses NEC custom ASIC for low latency and balance memory I/O design
- Azusa can be partition into four domains, each as an isolated and complete computer system
- An integrated service processor take care of platform management including preboot configuration, platform error handling, domain management, etc.
- Azusa is design with enterprise level RSA features for mission critical application

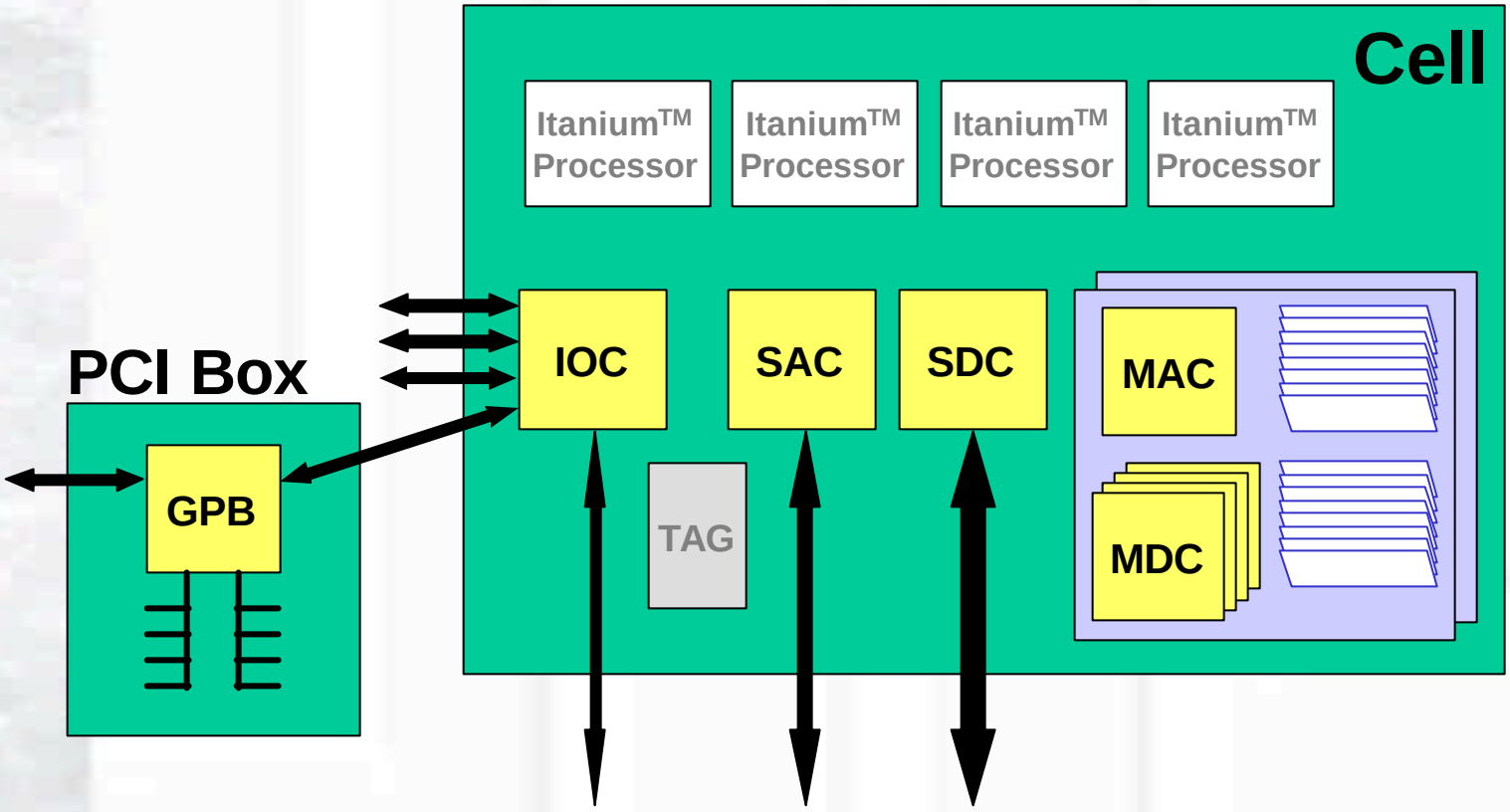
Azusa Overview



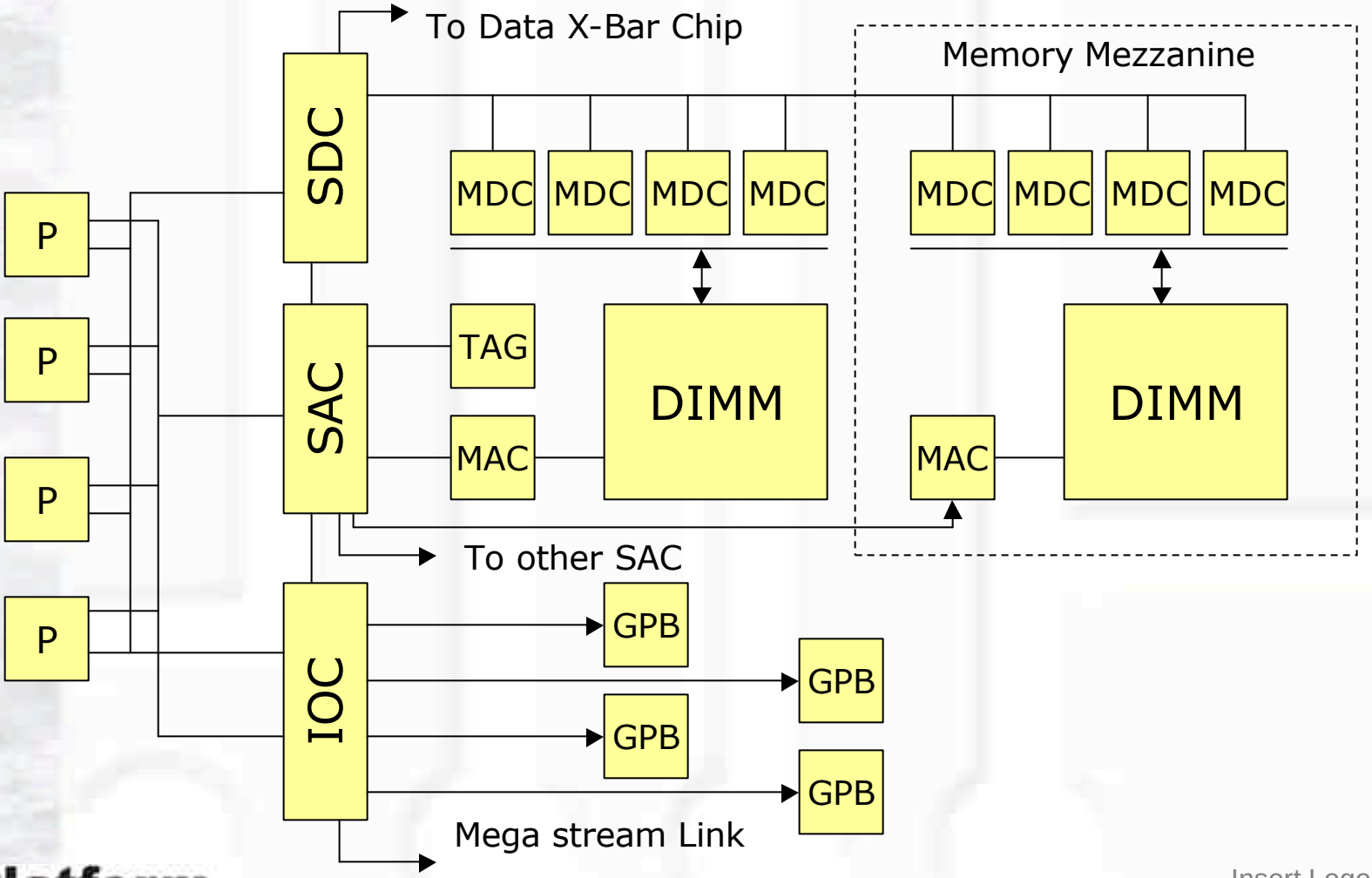
NEC Azusa Features

- Azusa connects four (4) Itanium CPU Quad cell using NEC Custom ASIC and high speed Cross bar switch to gain non-blocking access to remote memory
- CPU Quad cell is hot swappable (isolated from X Bar)
- One (1) service processor (SP) connect to base I/O ports for system functions
- Each Quad cell can hold 32 GB of memory (on 32 DIMM) for total of 128 GB memory in full 16 way configuration
- Azusa can be configured as one (1) system with 16 CPU or four (4) domains with four (different) OS instances

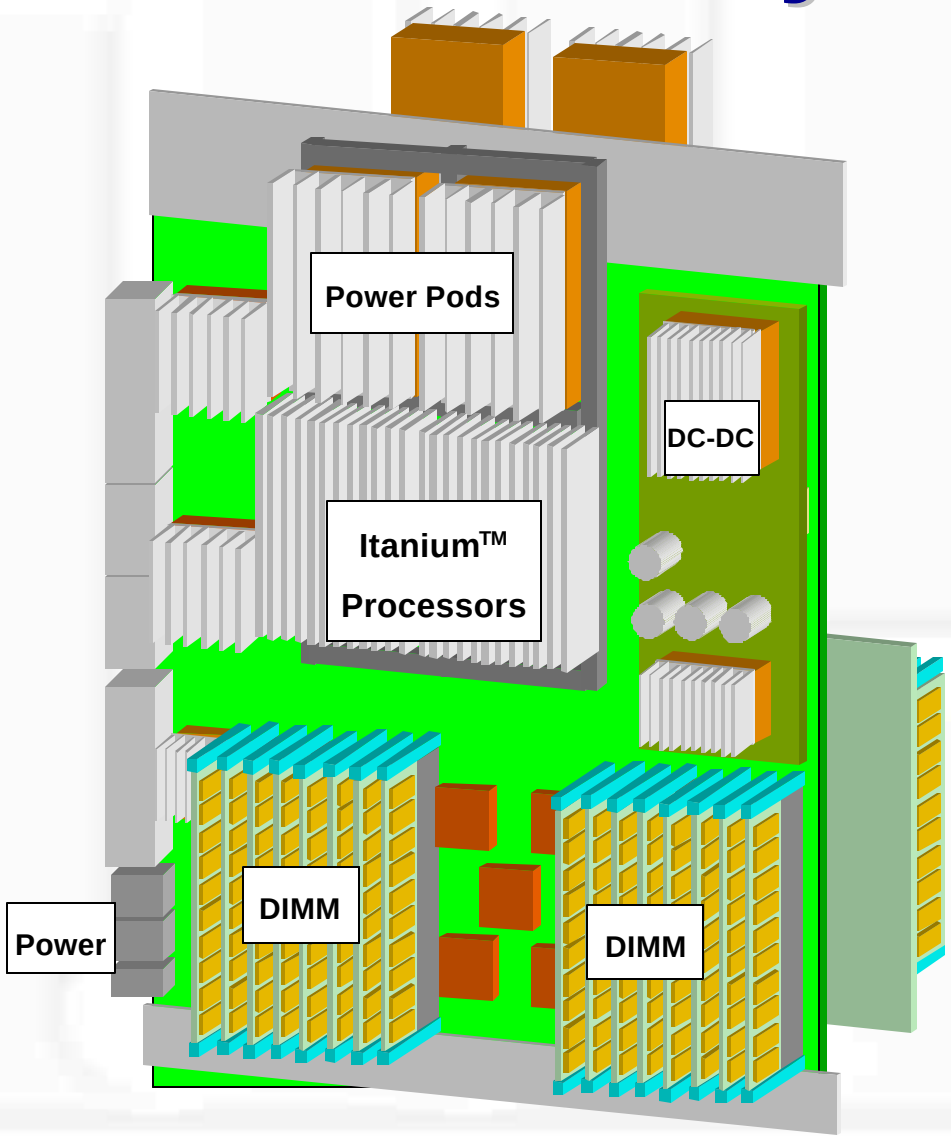
Cell & Chipset



Azusa Chipset



Azusa Assembly

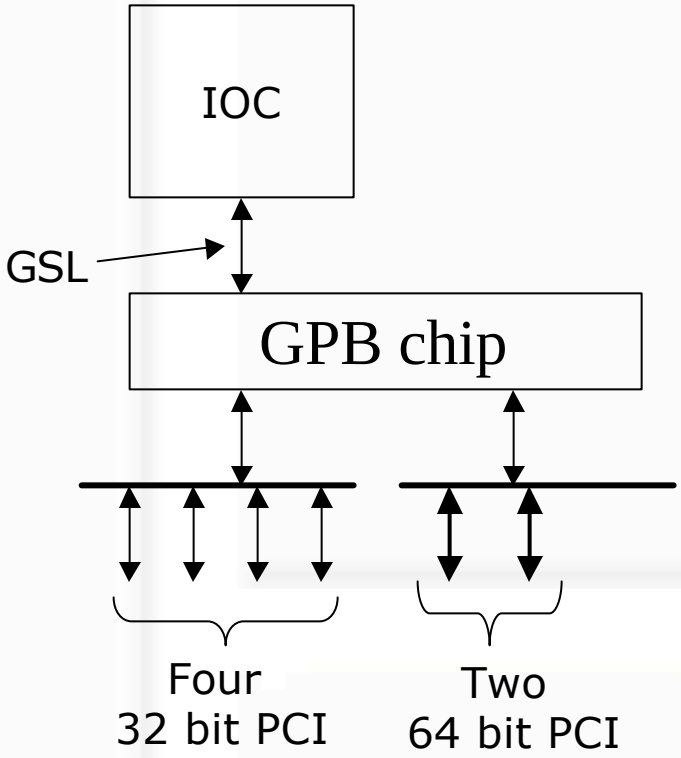


Azusa IO Structure

- Azusa has PCI in each cell; each PCI adapter has two 64-bit PCI buses that are configurable as either two slot 66 MHz buses or four slot 33 MHz buses
- All PCI slots are hot plug-able
- Each PCI adapter has 2 GSL port; can be used together for performance or alternatively for redundancy
- Maximum is 128 PCI slots or 32 buses (on 16 PCI adapter total) for 8 buses per each Quad cell
- Total I/O bandwidth is 8 Gbytes/sec

GSL

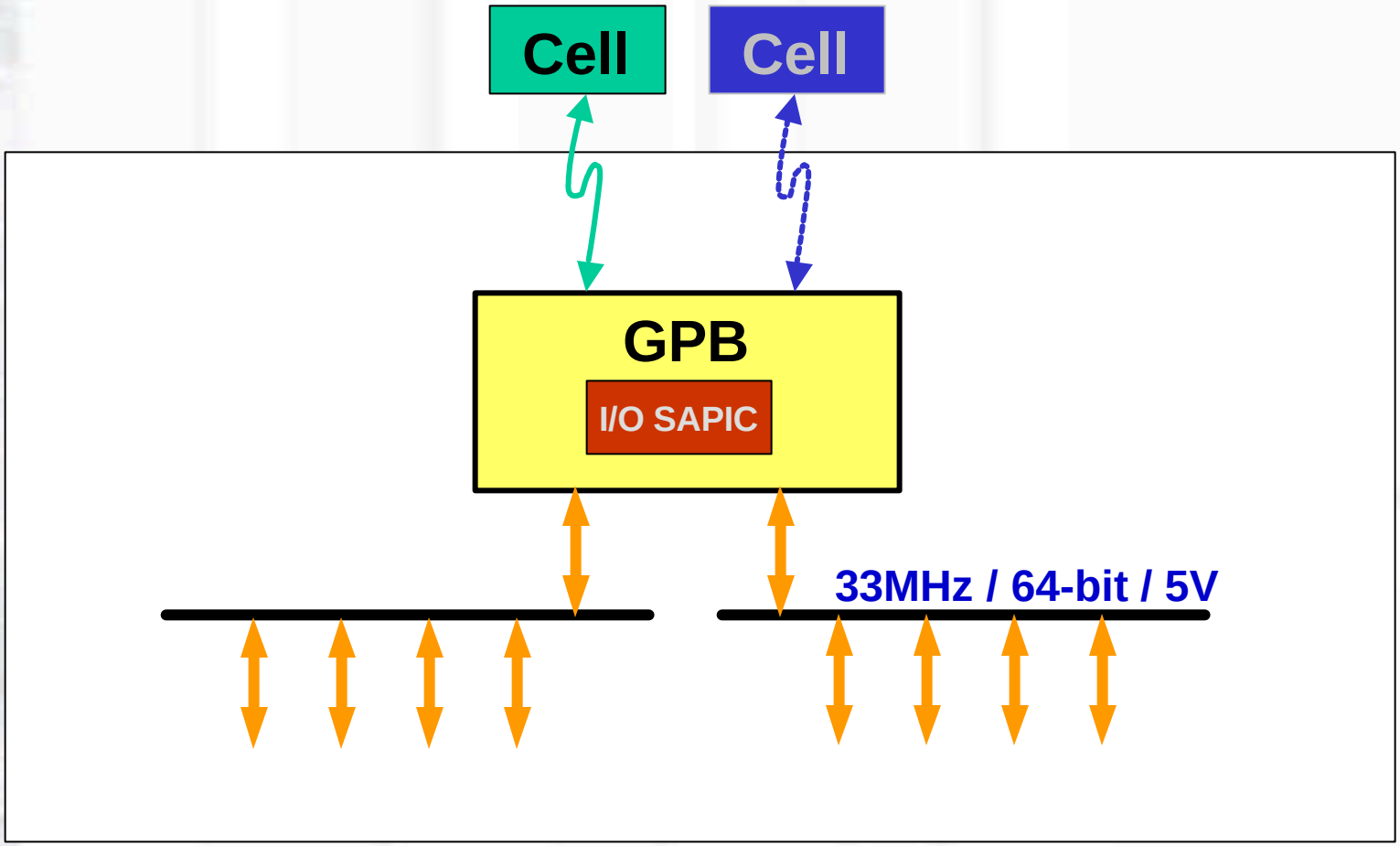
- Giga-stream Link on copper technology
- Max length is 5 M (16 feet)
- GSL is used to connect from CPU Quad cell to PCI adapter to break out buses
- GPB = GSL to PCI Bridge
- I/O Streamlined advanced programmable interrupt controller (Sapic) inside GPB



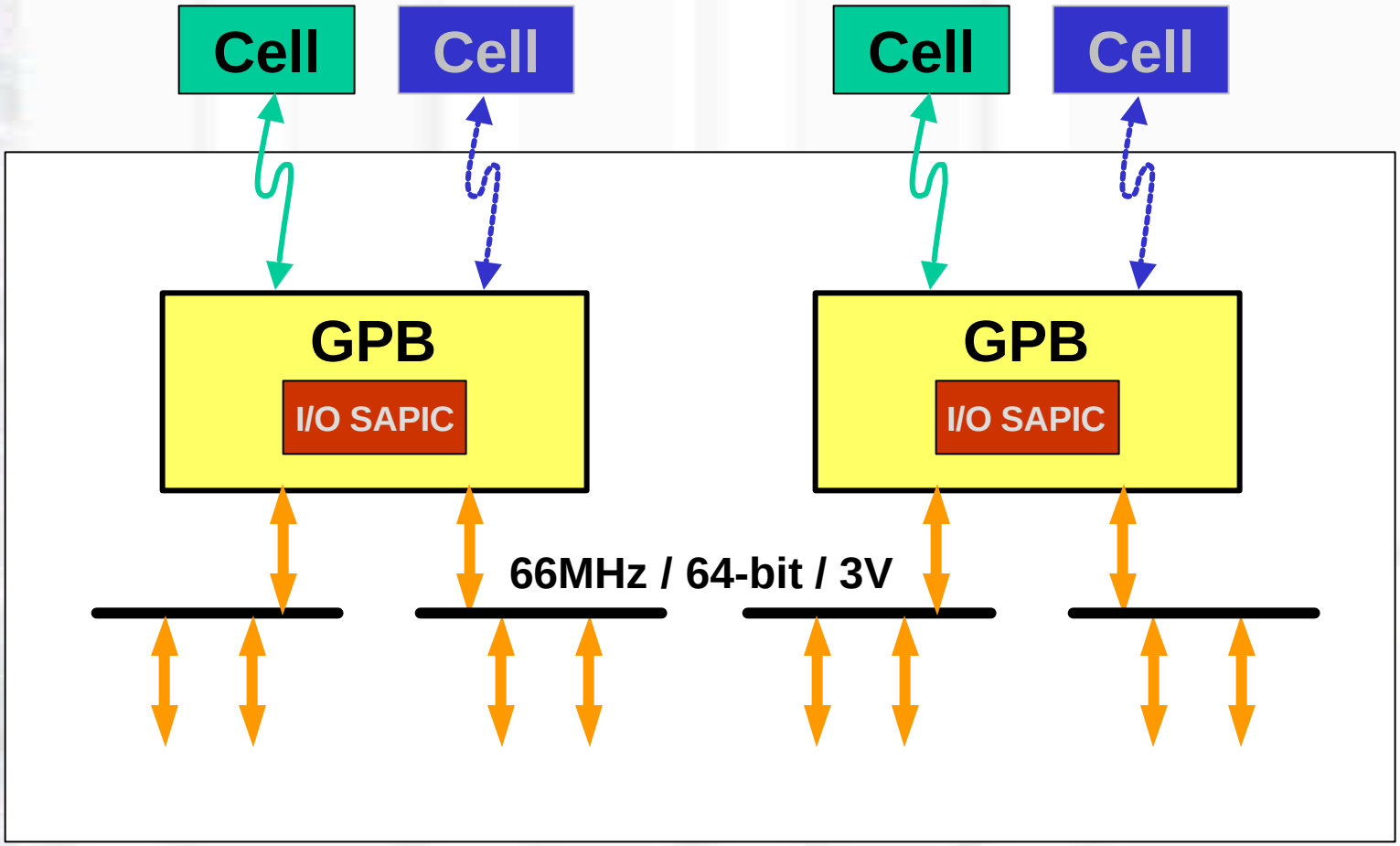
Azusa I/O and GSL

- GSL can support both 66 MHz and 33 MHz and operate independently
- PCI-X version of GPB will be ready in time
- GPB perform write combine for subline inbound stores and prefetch for better DMA read
- GPB can also perform peer-to-peer transaction as well as legacy sideband signals mode
- GPB contains SAPIC which can also support 8259 legacy modes

PCI Box - 33MHz



PCI Box - 66MHz



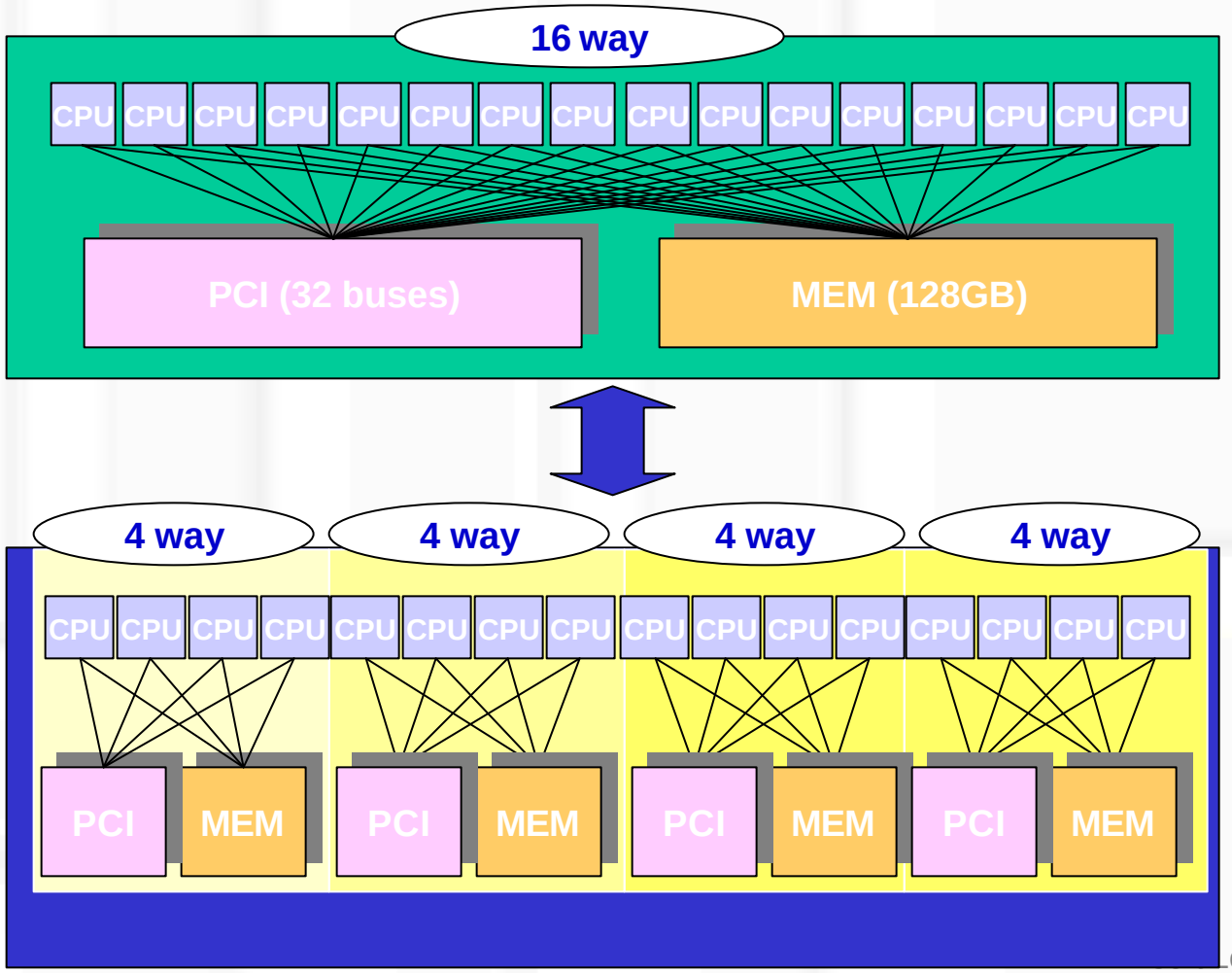
SAC, SDC, and IOC

- SAC handles system bus, I/O and inter-cell transaction; internal/external coherency control, address routing, etc.
- SAC also control SDC and how data flow in and out of the SDC to memory, IO and CPU
- Each memory bank consists 1 MAC and 4 MDC
- Azusa supports chip-kill features, memory scan, dynamic test at power on and periodic memory patrol and scrubbing
- IOC Mega-stream link connects to the legacy south bridge and service processor
- IO transaction look aside buffers are integrated in the IOC and converts a 32 bit address issued by a single address cycle PCI device into a full 64 bit address

Azusa ccNUMA

- Azusa memory latency characteristics is very close to large SMP and NUMA ratio is 1.5 (best case)
- SAC contains Tag SRAM to reduce snoop traffic that keep tracks of all 4 CPUs; when a coherent memory transaction is issued in one cell, its address is broadcast to all other cells for simultaneous snooping.
- Snoop filter will compare the address, if hit, then it will forward to system bus for snooping; if miss, then the overhead is only tag access
- Either way snoop filter is updated by replacing or purging the tag entry associated with the CPU cache line that was loaded with the memory data

AzusA: Software Perspective

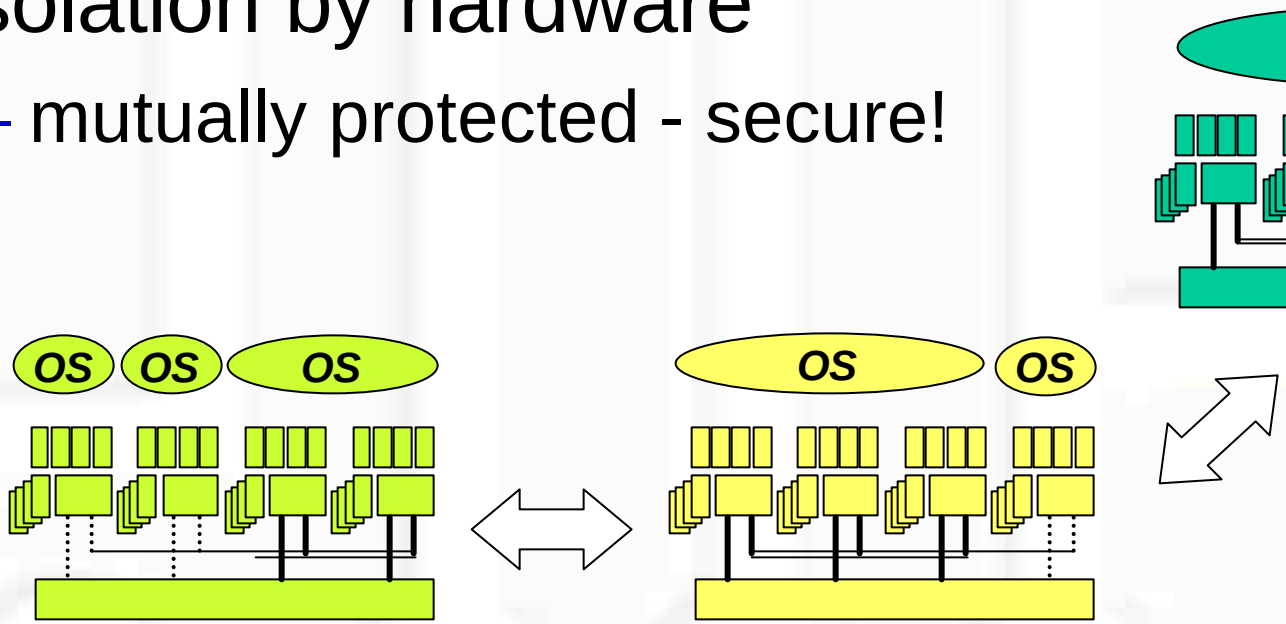


Azusa ccNUMA

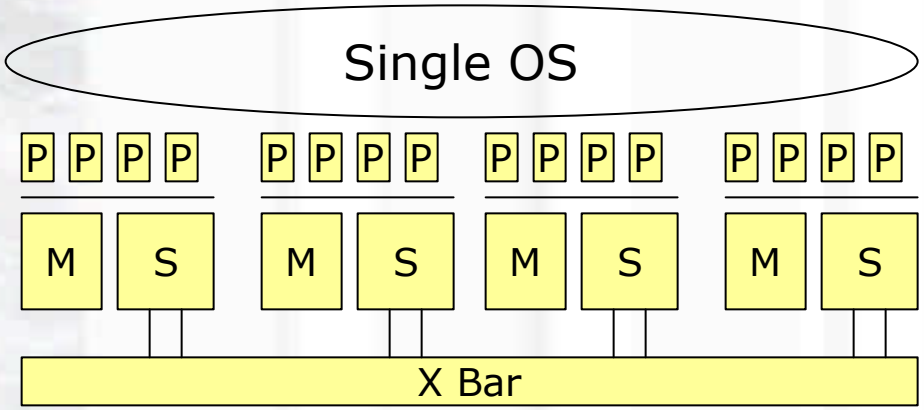
- Memory read is speculative whether the line memory is cached or not
- Each cell has about 4.2 Gbytes/sec bandwidth (total 16.8 Gbytes/sec)
- SDC to data cross bar is 8.4 Gbytes/sec total for full performance
- SAC use memory range register to configure memory to present to OS as one single linear memory space
- All PCI can be also configured as single PCI tree during boot time or on line
- Configuration register are mapped as Intel 82460GX chipset so it can be an extension

Hard Partitioning

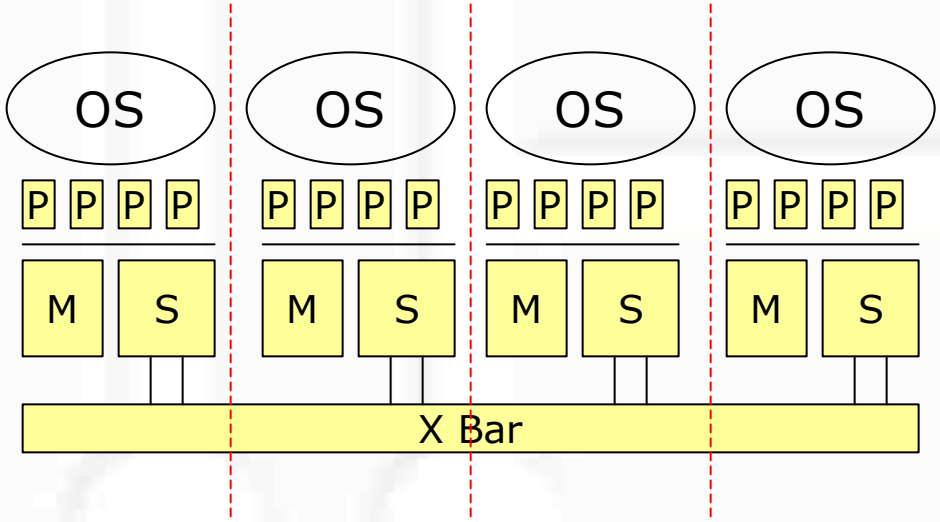
- Arbitrary number of cells
- Arbitrary number of domains
- Isolation by hardware
  - mutually protected - secure!



Azusa Configure Mode



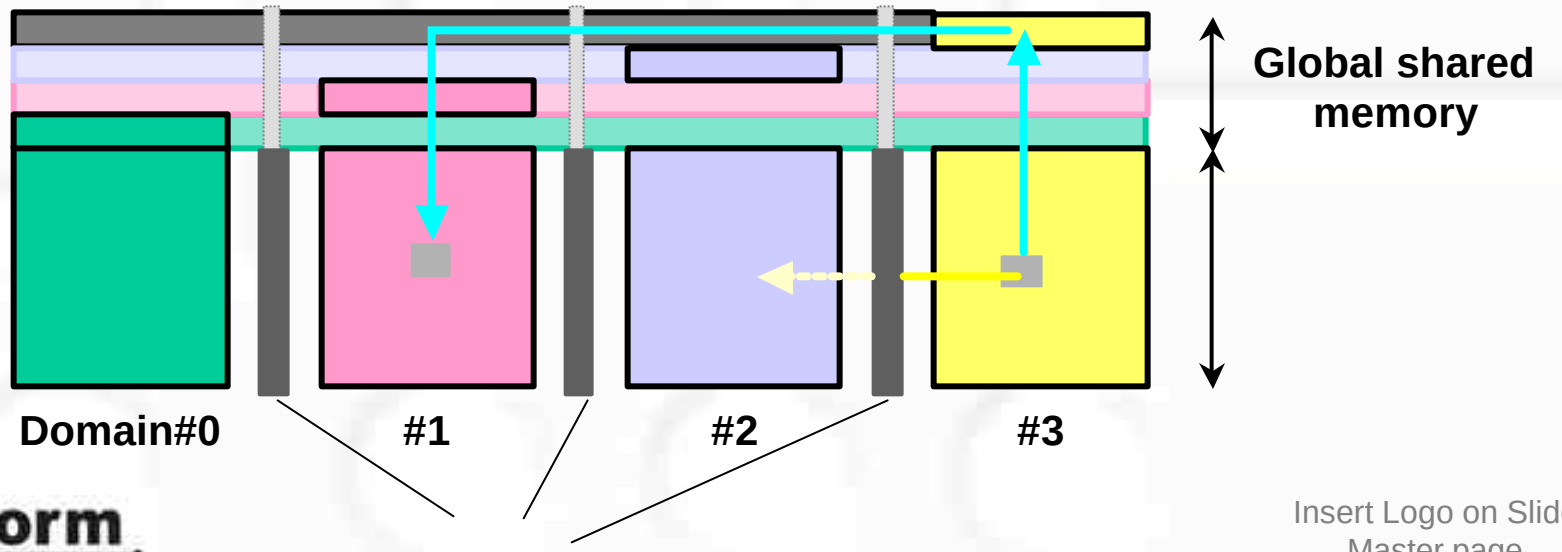
Single OS, 16 way SMP machine for large database or transaction operation



Four OS, 4 way SMP machine each, with hard protection between each domain

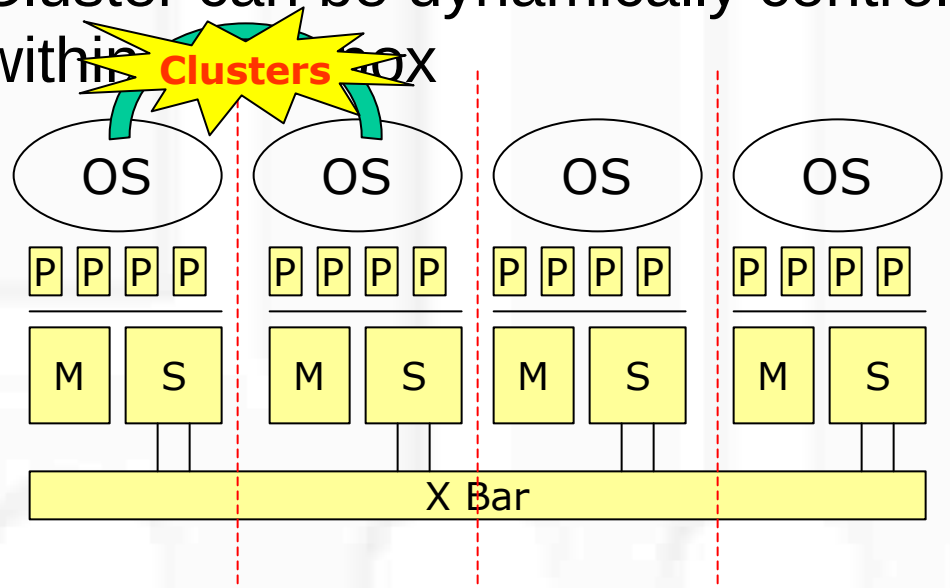
Inter-domain Communications

- Communications via physical memory
  - Configurable globally shared memory
- Standard API support
  - VIA and others



In Box Cluster

- When partition as multiple domains, Azusa provides in box cluster to provide HA (High Availability) features
- Cluster can be dynamically controlled and managed within box



Can run as independent box and join in cluster later if necessary

Allow ASP to sell modulized service

Standards Conformance

- AzusA conforms to *DIG-64* and other industry standards
  - DIG-64 : *Developer's Interface Guide for IA-64 Servers*
  - EFI : *Extensible Firmware Interface Specification*
  - IPMI : *Intelligent Platform Management Interface Specification*
  - HDG : *Hardware Design Guide for Microsoft Windows 2000 Server*
  - ACPI : *Advanced Configuration and Power Interface Specification*
  - SSI : *Server System Infrastructure Specifications*

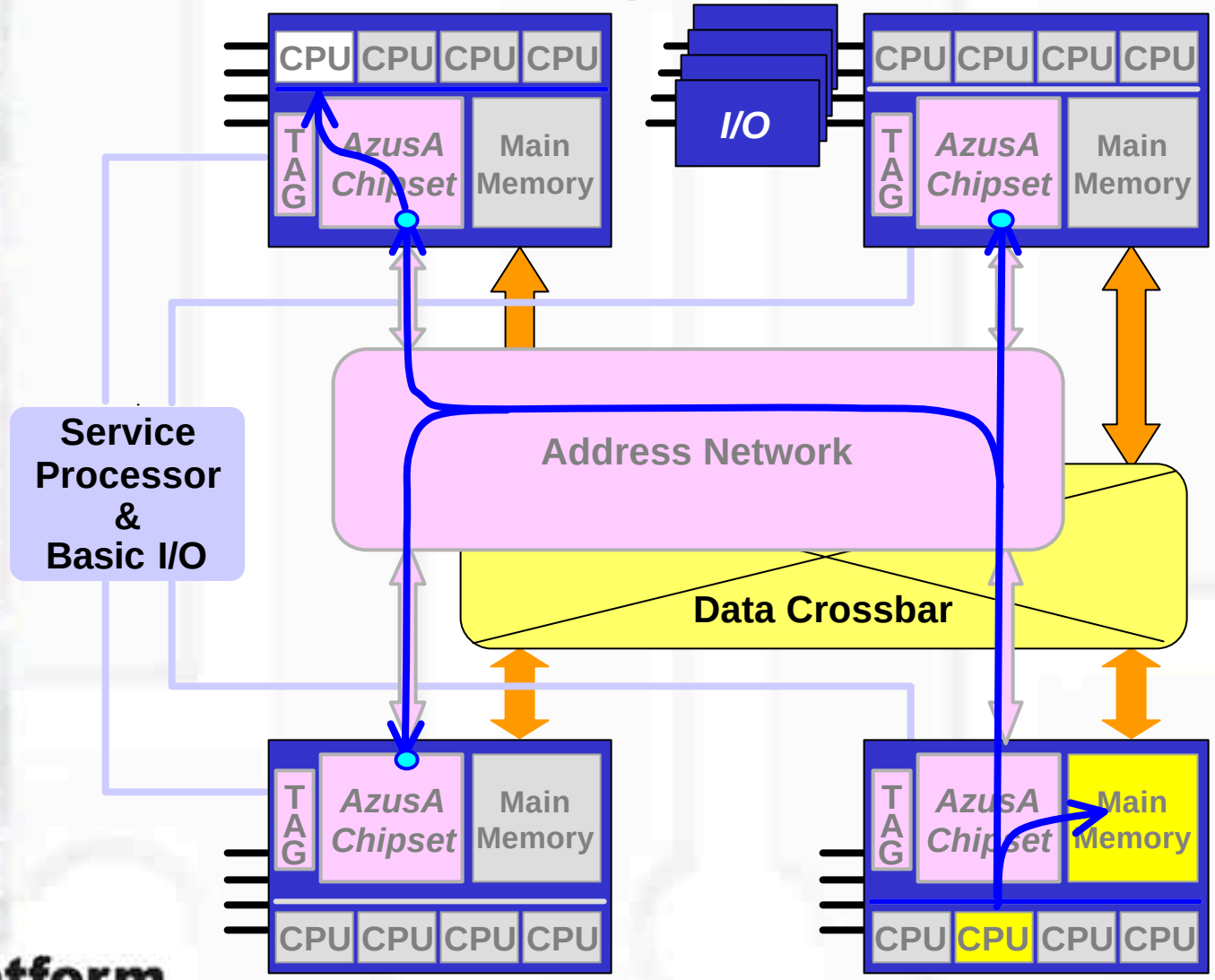
Low Memory Latencies

- Local latency: < 200 nanoseconds
- Remote latency: < 300 nanoseconds
- Low I/O latencies
- Low cache-to-cache latencies

Snoop-based Coherency

- Fast snoop via address network
- Snoop filters to reduce remote snoop traffic to CPU bus
- Chosen over directory-based scheme
- Performance and cost advantages
  - Very low cache-to-cache latency
  - Lower average latency
  - Cost not proportional to memory capacity

Snooping in AzusA

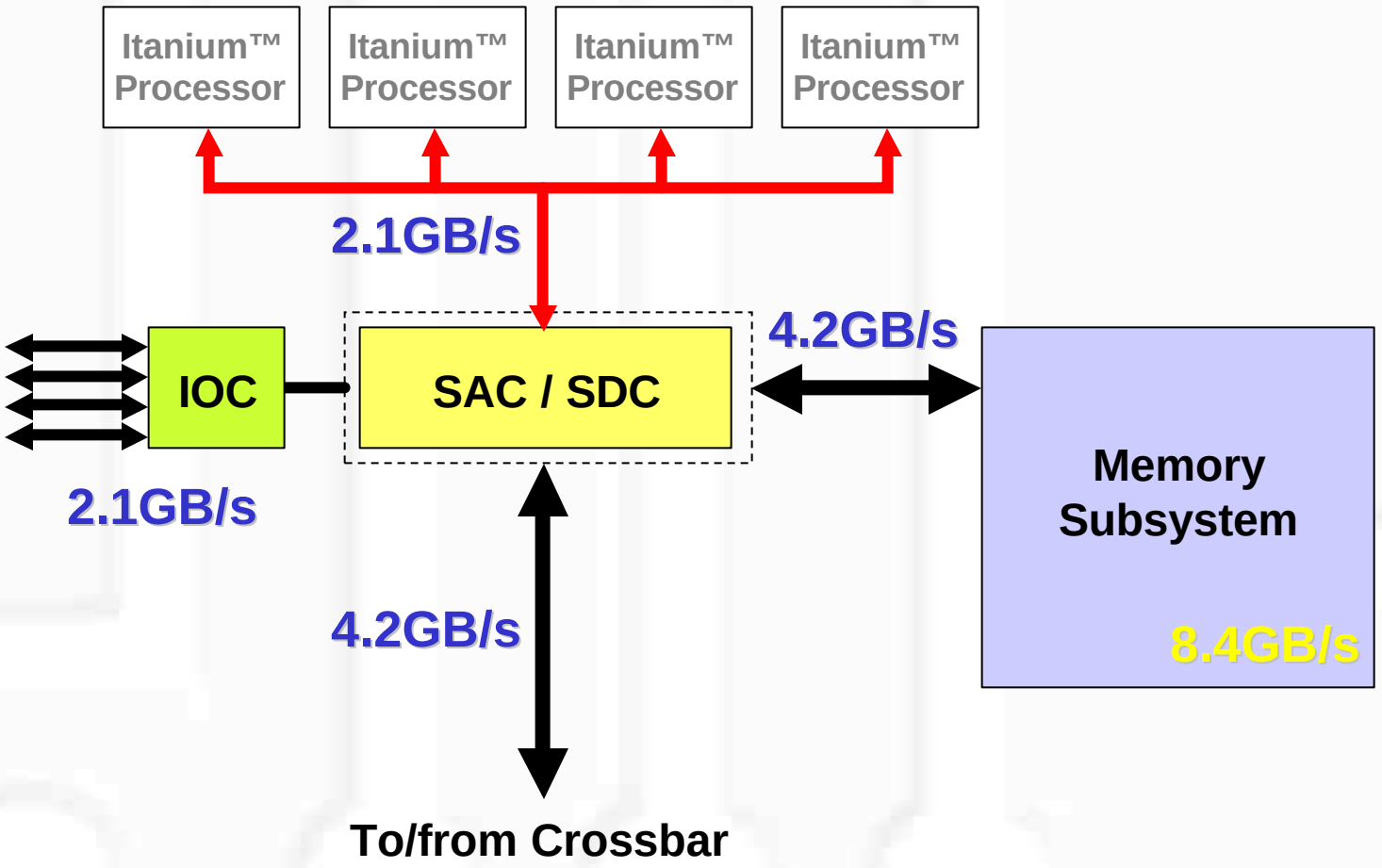


Balanced and High Bandwidth

- Memory can sustain > 2x CPU Bus data bandwidth
- Cell memory bandwidth can be fully exported to other cells via crossbar
- Snoop bandwidth > data bandwidth
- Balanced I/O bandwidth

High Bandwidth = High Scalability

Balanced Bandwidth



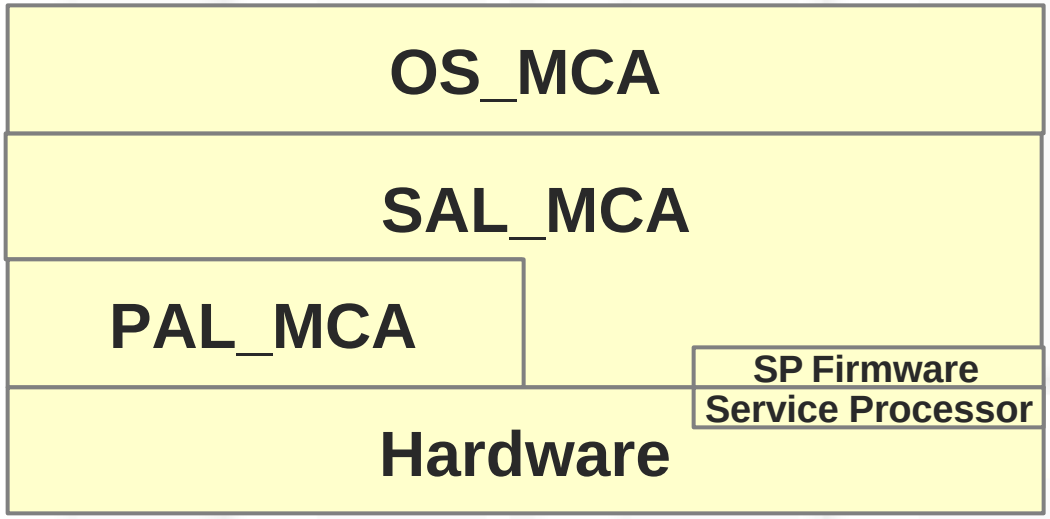
Azusa RSA

- Modules are Hot plug-able
- Memory mirroring, data path protection and correction, error containment
- Graceful error propagation
- IA 64 machine check abort recovery
- Parity protected control logic
- Hardware consistancy checking
- Detail log registers
- Service Management access to chipset

MCA Architecture

- Integral part of IA-64 architecture
- Layered architecture
- Defines error handling framework
  - logging and reporting
  - foundation to advanced error recovery
- System could survive ‘fatal’ errors
  - by aborting affected process only
  - with platform error containment and careful OS_MCA recovery

MCA Layers


OS_MCA
SAL_MCA
PAL_MCA
SP Firmware
Service Processor
Hardware

Chipset RAS Features

- MCA recovery support
  - error containment and precise reporting
- Parity/ECC protected control logic
  - not just arrays
- Cell hot plug support
- Extensive log information
- Back door access port
- Watchdog timer

Memory RAS Features

- Error containment with data poisoning
  - enables MCA recovery from uncorrectable ECC errors
- *"chip kill"* configuration
- Patrol & scrub
- Dynamic memory de-allocation
  - preventive re-configuration on the fly

Cabinets

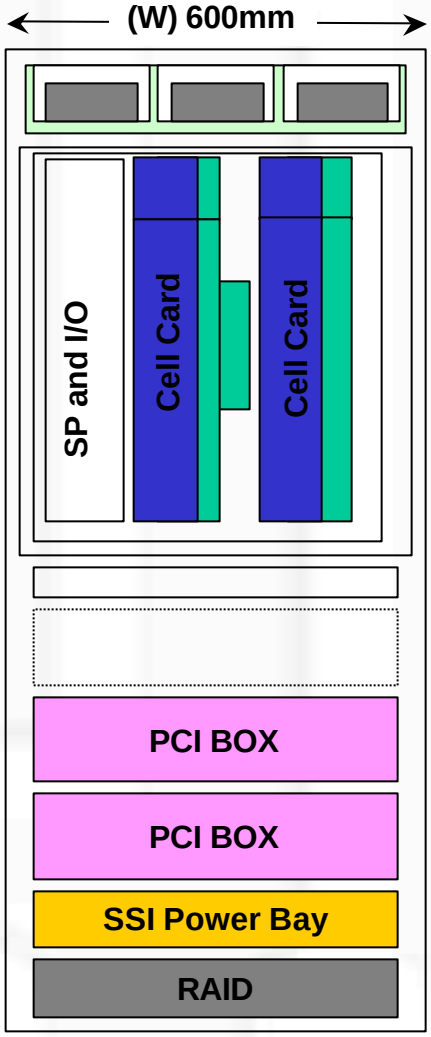
Expansion

Main

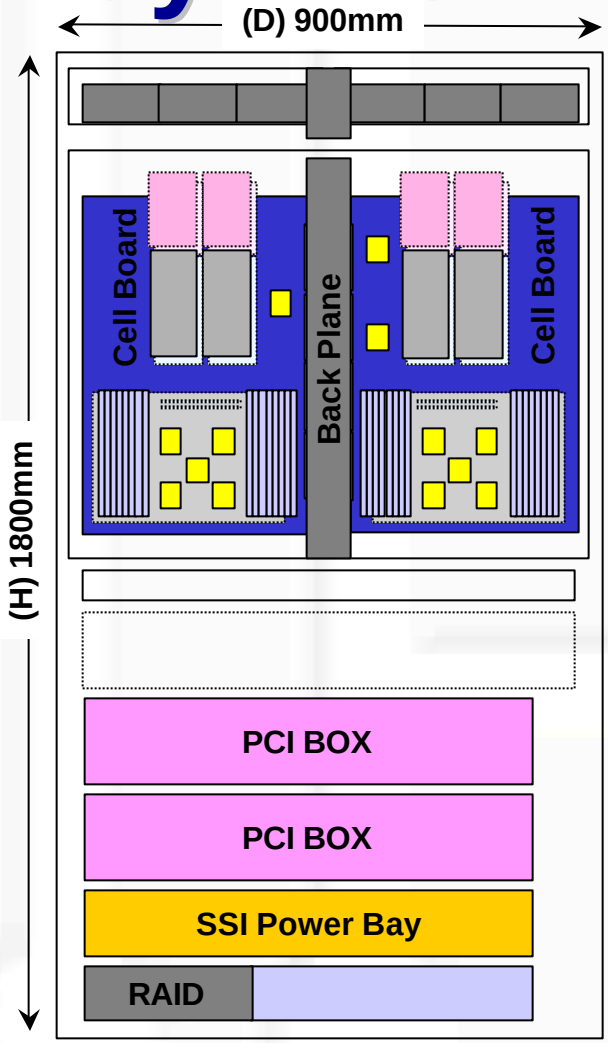
Platform Conference

Logo on Slide
aster page

Phvsical Layouts



Front View



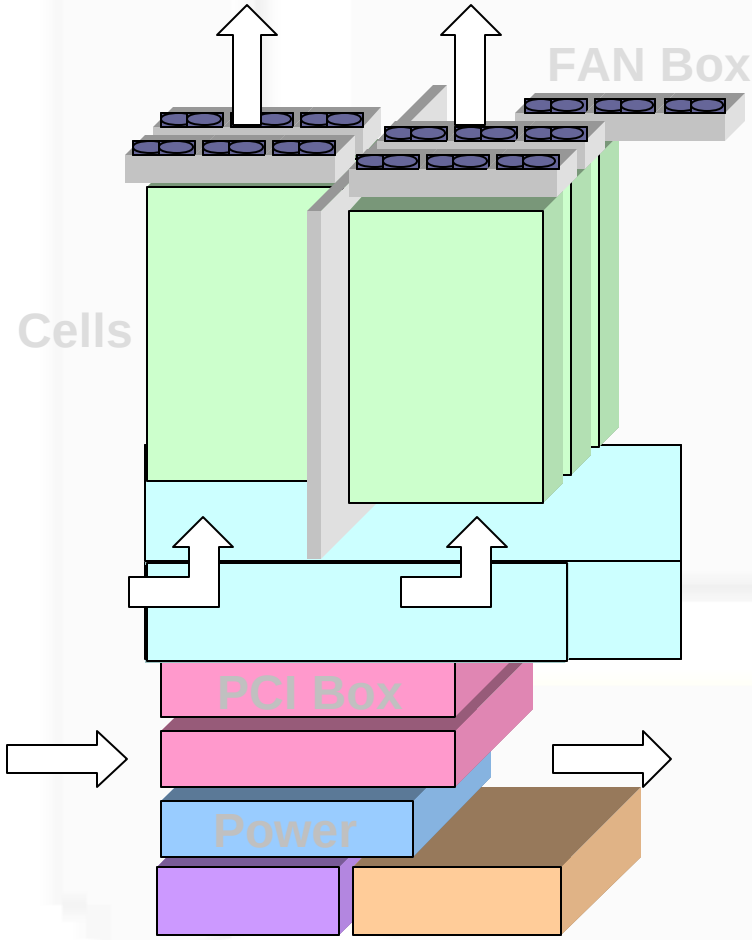
Side View

Platform Conference

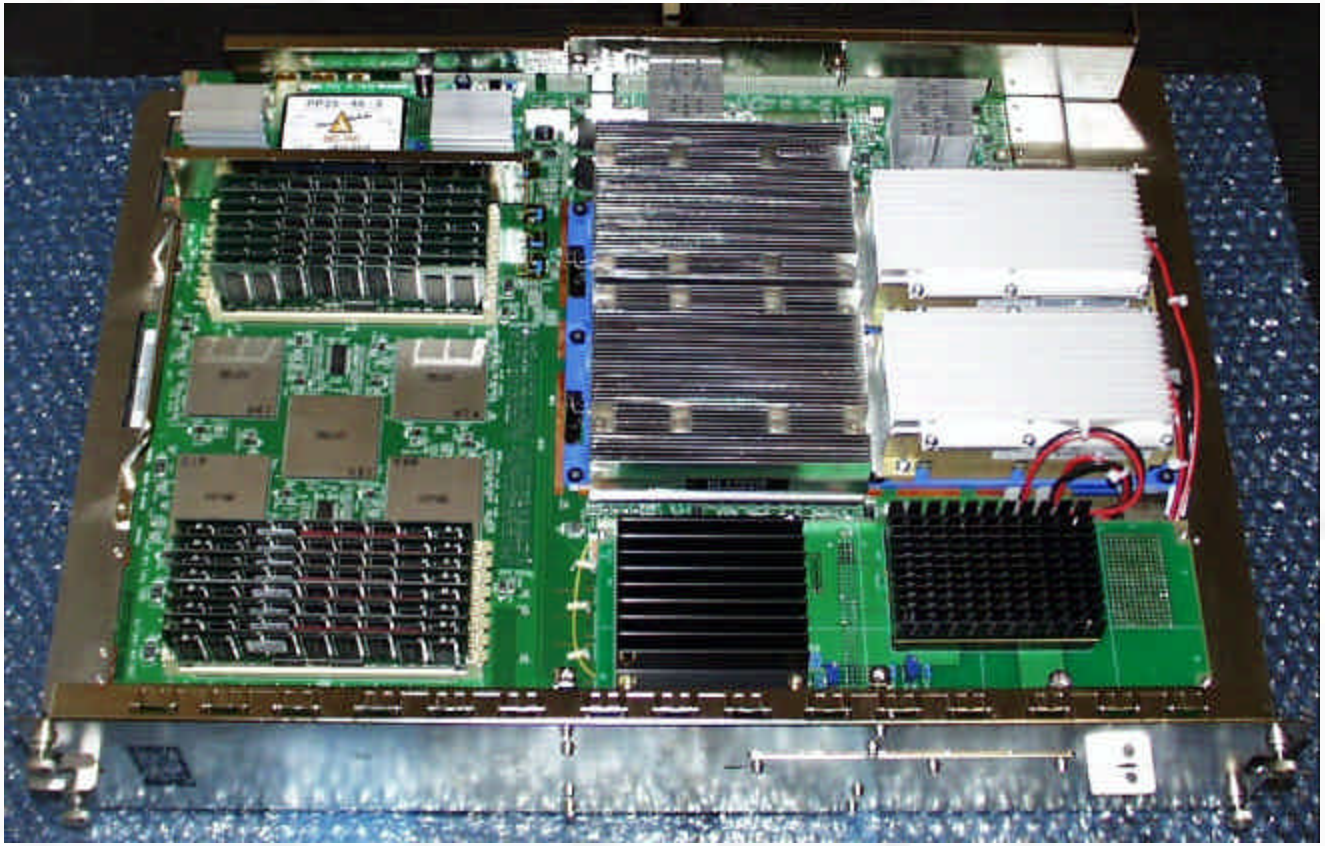
Insert Logo on Slide Master page

Cooling and Air Flow

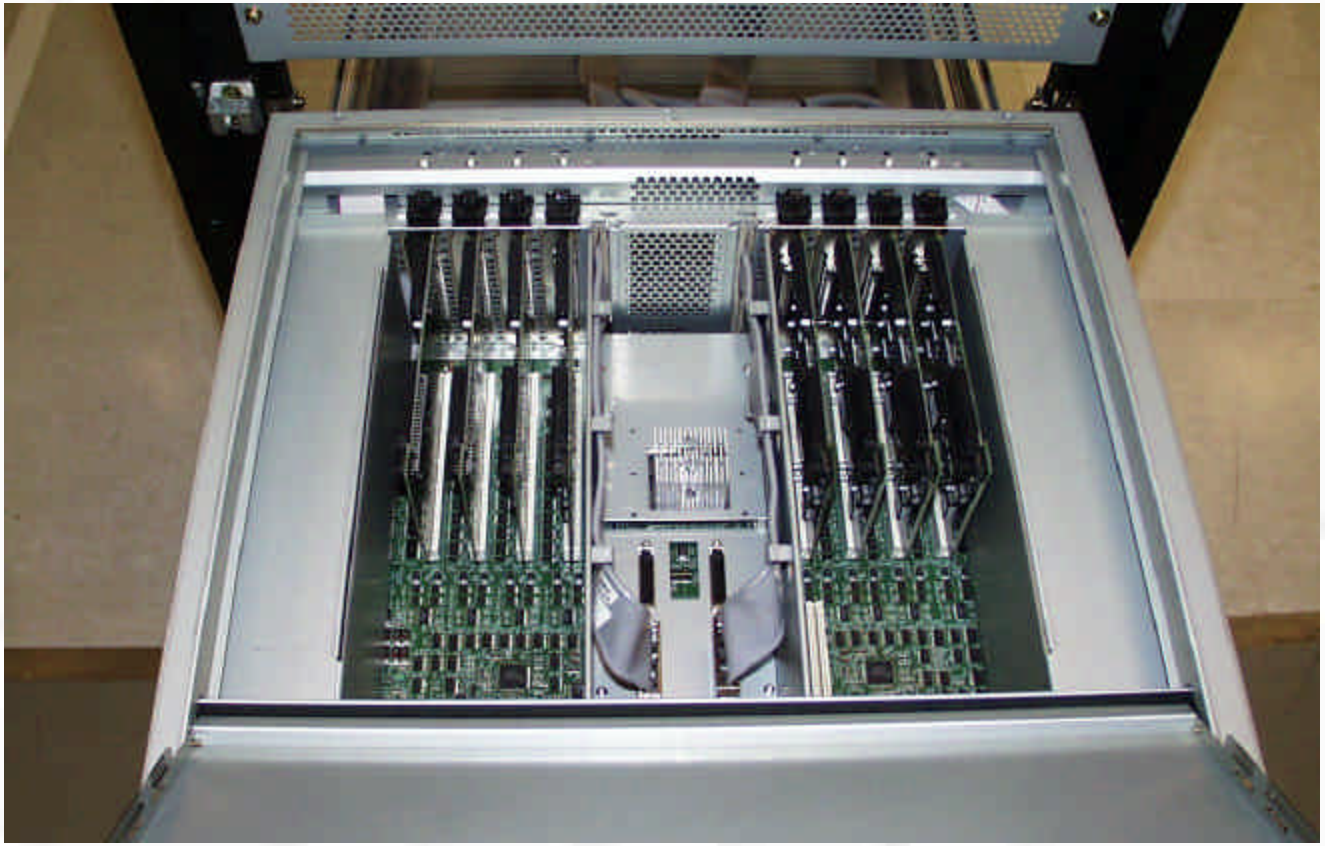
- Hot swappable fans
  - redundant
  - multi-speed
  - low noise



Prototype Cell



Prototype PCI Box



Outline

- Path to HPC
- NEC Azusa overview
- **Challenges**
- Future
- Q & A

ILP, TLP and SMP

- instruction scheduling and predicate helps ILP
- CPU affinity helps TLP – OS lacks the sophistication
- Single OS imaging create unnecessary remote access; true micro kernel may help
- Software pipelining, loop unrolling and code explosion has different effect on different size of SMP, NUMA
- Compiler and dev environment needs major retooling
  - Unfortunately, this takes about 10 years
- Big Bus SMP does not work – snoop traffic problem

OS consideration

- No off the shelf OS now is NUMA aware
  - Vendor are require to do special OS support to gain full access to the HW and feature set
- OS expect single memory, single CPU
  - HW design has to conform to this and limit the NUMA flexibility and design approach
  - Future distributed computing needs to break this
- Future OS needs to have concept of memory access “cost” (NUMA ratio) and “locality” (remote vs local)

OS and API

- OS needs to provide transparent memory management to schedule resource accordingly
  - Programmer can not assure what system the software will run on; even EPIC can't help this
- Red Zone and MM design mismatch may cause hot spot which stall NUMA and have to be avoided
  - Modulize OS kernel and function repartition is necessary for future larger NUMA system where cc breaks down
  - How do we discriminate memory in current OS?

Physics and Reality

- High speed design, package and density
  - Bus loading and flight time; point to point with pin count and 3D trace routing; clock distribution and synchronization
  - Complex system and circuit level simulation
  - Proprietary and specialized components and design
- Thermal and Power
  - Limit design flexibility and long term reliability
- Economy – Infrastress factor
  - Market size and sales support

Interconnect

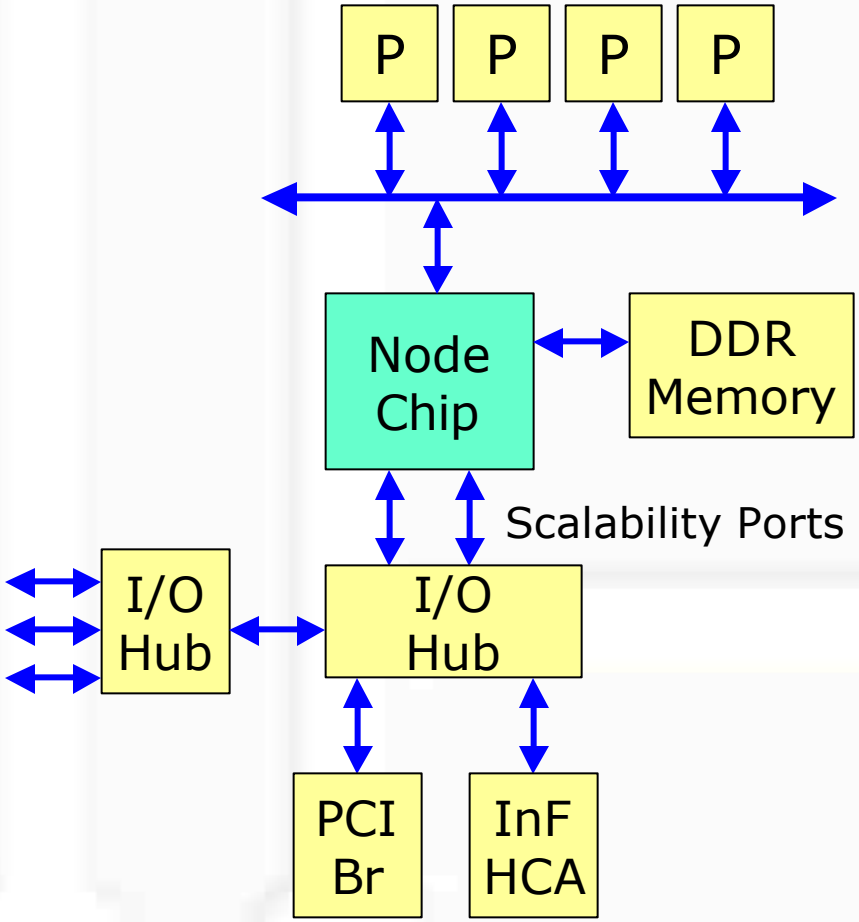
- SCI and HIPPI
  - Need special ASIC and cost a lot more for ccNUMA
- NUMALink, XIO
  - SGI inherit from Cray, used in NUMAFlex Origin series
- Infiniband
  - Intel will drive the market and ASIC development
  - Prototype demo at IDF and (almost) ready for market now
- LDT
  - Please see AMD presentation for detail

Outline

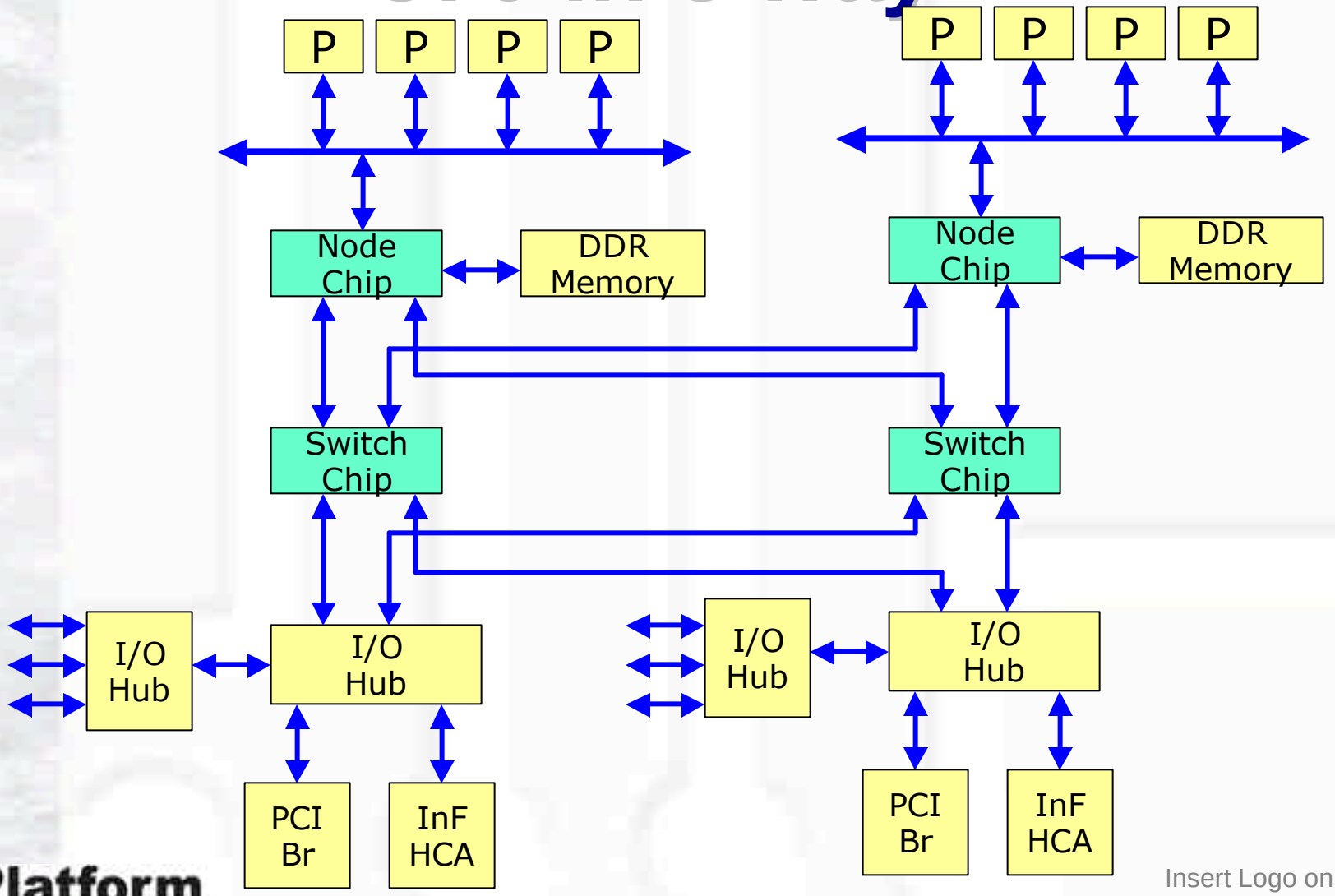
- Path to HPC
- NEC Azusa overview
- Challenges
- **Future**
- Q & A

870 & Scalability port

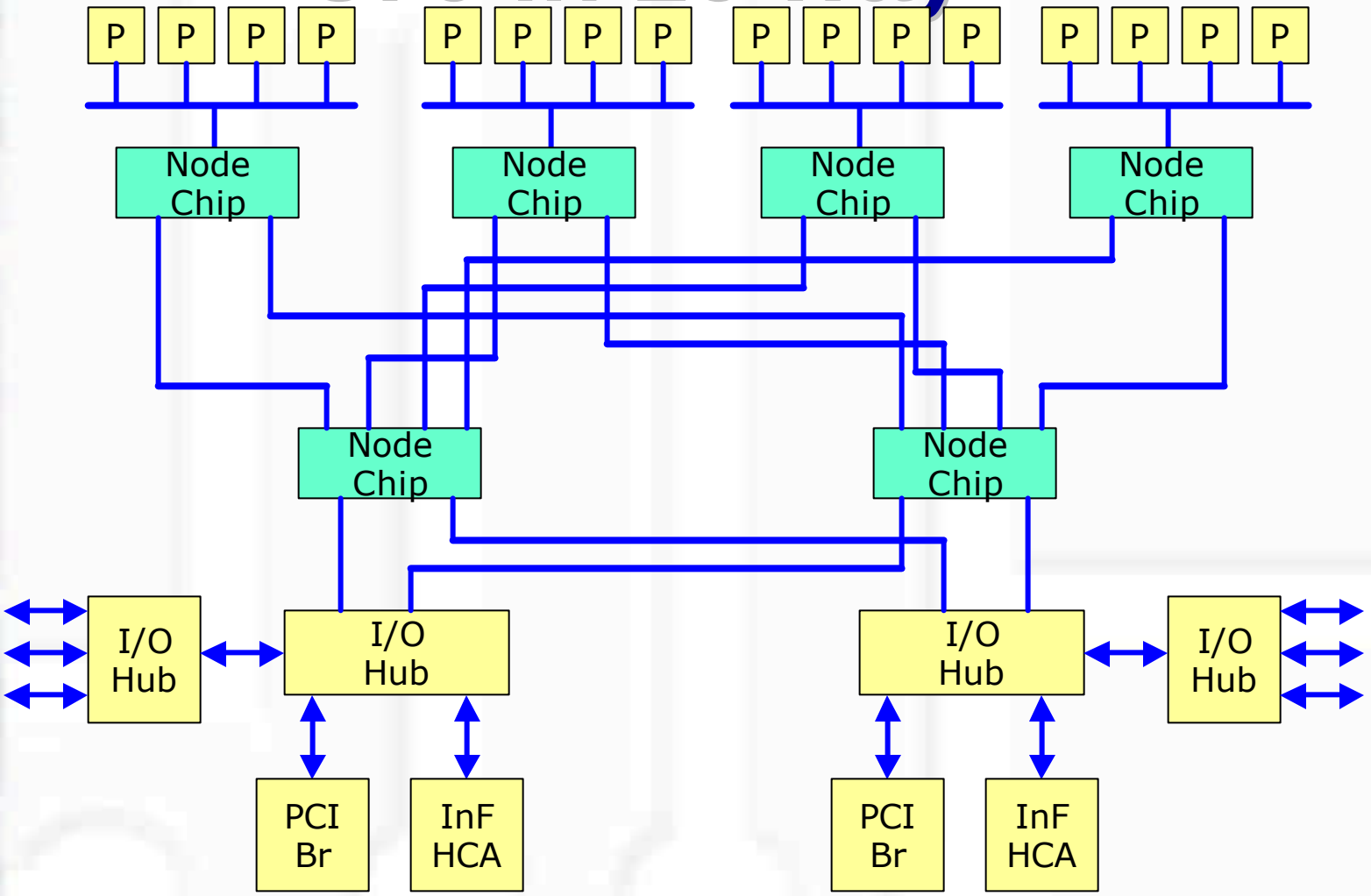
- Serves McKinley and future IA64 CPU
- Node chip uses switching architecture to balance port, memory & I/O request
- Both server and WS



870 in 8 way



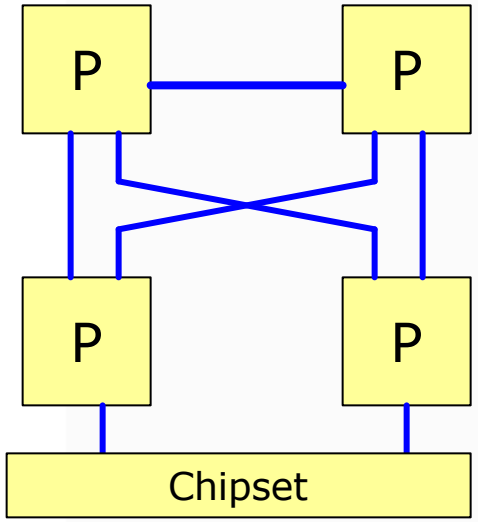
870 in 16 way

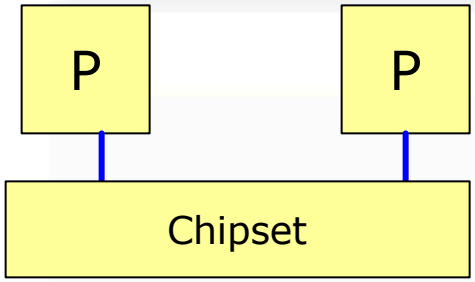

P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
Node Chip
Node Chip
Node Chip
Node Chip
Node Chip
Node Chip
I/O Hub
I/O Hub
I/O Hub
I/O Hub
PCI Br
InF HCA
PCI Br
InF HCA

X86-64 & LDT

- 2 way, 4 way and possible 8 way using LDT
- Direct CPU to CPU link at very high BW
- Provides snoop, IPC and other I/F in packet format

- Detail restricted under NDA





Extend the future

- Moore's Law
  - Processor and Micro Architecture improvement – move up to McKinley and "future" IA64 CPU
- Extend to larger SMP
  - Support 32 way in dual rack configuration with larger RAM and a Admin console server
  - Exposed Clustering in data center to MPP size
- True NUMA OS, compiler and Software
  - New OS for distributed computing from design

Base line support

- Support IA 64 machine check architecture in OS
  - OS HAL does not take advantage of HW features
  - No graceful degradation of PU and memory – this makes it hard to reach 5-9 reliability target
- NUMA aware OS
  - Distributed, micro-kernel with localized service layer
  - Machine specific OS porting support Asymmetric and dynamic configuration

Intelligent Installation

- Defer optimization and code generation to install time
  - Escrowed cabinet deliver secure obj files
  - OS feedback actual system setup at run time
  - Installer invoke linker and optimization and perform code generation
  - Machine specific code is optimized per each system depends on available resource
  - Low programmer impact where most of intelligence is in installer; OS vendor can standardize this

Multi-level Clustering

- From Soc MP to MPP cluster
  - Support multi-level clustering
- Cascading resource
  - Allow Resource pooling at PU level
  - Nanoclustering and Wafer scale integration in both structure and non-structure dynamic network
- HW Directory service
  - E.g. HP System Locality Information Table
- Dynamic Partitioning

Outline

- Path to HPC
- NEC Azusa overview
- Challenges
- Future
- **Q & A**

Question ?

Reference

- F. Aono, M. Kimura“The Azusa 16-way Itanium server”, IEEE Micro Sep/Oct 2000 pp54-60
- Gregory F Pfister, In Search of Clusters, 2nd ed., Prentice Hall PTR, 1988
- Advanced Configuration and Power Interface, http://wwww.teleport.com/~acpi
- Intelligent Platform Management Interface, http://developer.intel.com/design/servers/ipmi/index.htm
- Server System Infrastructure, http://www.ssiforum.org/default.asp
- The Developer’s interface Guide for IA-64 Servers, http://www.dig64.org
- AMD x86-64™ Architecture Programmers Overview, http://www.amd.com/products/cpg/64bit/pdf/x86-64_overview.pdf
- HP Itanium Resource page http://devresource.hp.com/devresource/Topics/IA64/IA64.html
- P. Markstein, IA-64 and Elementary functions, HP professional books

Reference

- I. Foster, GRID, Morgan Kaufmann Publishing
- Peer to Peer working group spec, http://www.peer-to-peerwg.org/specs_docs/index.html
- Microsoft 64-Bit Platform Design web site, http://www.microsoft.com/HWDEV/64bitWindows/default.htm
- 64bit Itanium linux beta from redhat.org, ftp://ftp.redhat.com/pub/redhat/ia64/
- A. Reinefeld, SCI: Scalable Coherent Interface, Springer Verlag
- 1596-1992 : IEEE Standard for Scalable Coherent Interface, Sci, ANSI
- Hot Chip 12 conference proceeding
- 2000 Microprocessor Forum handout